[주제 1]

집단 비교 분석 방법

2018년

변종석(jsbyun@hs.ac.kr)

집단 비교 방법의 이해

1. 분석 방법의 기본 개념
2. 모수방법
   1. 두 독립 집단의 비교 : T-검정
   2. 대응 자료의 집단 비교 : 대응 T-검정
   3. 세 집단 이상의 비교 : ANOVA
3. 비모수 방법
   1. 독립집단의 비교
   2. 대응집단의 비교

서론 : 개요

- 생명과학 연구란?
  - 의학, 약학, 간호학, 보건학, 역학 등 생명과학 관련 분야에서 연구 계획과 설계, 자료 수집, 분석 및 해석 등 일련의 과정을 진행하는 연구

- 생물통계학(Biostatistics)
  - 의학 등 생명과학에 대한 기초 지식을 바탕으로 생명과학 연구의 절차나 과정, 연구자료의 특성을 이해하고, 연구를 위한 설계, 자료수집, 분석 및 해석 등에 대한 방법론을 다루는 통계학의 한 분야
  - 자료 특성 : 사건(event), 생존시간(survival time), 절단자료(censored data),
  - 주된 분석 관점 : 비교연구, 생존함수 추정과 비교 등

- 주로 이용되는 통계분석 방법
  - 비교 연구 : two sample t-test, paired t-test,
    rank-sum test, signed-rank test
  - 연관성 분석 : Chi-square test, Fisher's exact test,
    McNemar's test,
    Cochran-Mantel-Haenzel test
  - 진단법 : sensitivity, specificity
  - 선형모형분석 : logit analysis, log-linear model
  - 분산분석 : ANOVA, ANCOVA,
    Repeated measure ANOVA
  - 생존분석 : Kaplan-Meier 방법, Cox-regression

1. 집단 비교분석을 위한 기본 개념

- 실험군과 대조군, 처치 전과 후의 비교가 주된 관점
  - 실험군(experimental group) : 어떤 실험의 대상이 되는 그룹
  - 대조군(control group) : 일반적인 상황에 노출된 그룹으로 실험군과 비교하기 위해 만든 그룹
- 가정 : 반응 결과가 연속형 자료로 수집

통계적 추론의 이해

- 통계적 추론 방법(Statistical inference)

  1) 추정(Estimation) :
     - 표본에서 얻어진 정보로부터 모집단 특성을 추론하는 방법
     - 점추정/구간추정 등으로 추정
     - 표본통계량, 신뢰구간, 신뢰수준이 중요한 도구

  2) 가설검정(Hypothesis Testing)
     - 모집단에 대한 가설을 설정한 후 표본자료를 통해 설정된 가설의 채택/기각여부로 추론하는 방법
     - 대부분의 통계분석은 가설검정 관점에서 분석
     - 가설, 검정통계량, 유의수준, 기각역 등이 중요한 도구

가설검정의 기본 원리

- 가설의 종류 : 귀무가설 H_0 v.s. 대립가설 H_1
  - 모수(parameter)에 대한 주장의 참, 거짓을 판단하는 과정
  - 연구가설 : 연구자가 주장하고자 하는 가설
  - 양측검정과 단측검정
- 가설 판단의 오류
  - 제 1종의 오류
    - 유의수준과 신뢰수준
  - 제 2종의 오류
    - 검정력
- 검정통계량과 표본분포
  - 검정통계량 : H_0 가 참이라는 조건에서 표본통계량을 표준화한 값
- 가설 판단 기준
  - p-value
- 결과 해석

가설 검정의 과정

① 가설의 설정

- 통계적 추론에서는 항상 2가지 유형의 가설로 설정

- 귀무가설(null hypothesis : H_0)
  - 기존의 이론적인 결과나 연구 결과로 설정
  - 차이가 없다는 주장으로 설정
  - 관계가 없다는 주장으로 설정
  - 예 : 기존 치료법과 새로운 치료법의 효과는 같다

- 대립가설(alternative hypothesis) : H_1 or H_A)
  - 귀무가설을 반박하는 대립적인 주장
  - 새로운 연구 결과로 설정
  - 차이가 있거나 크다/작다로 주장
  - 관계가 있거나 양/음의 관계 등으로 주장
  - 예 : 새로운 치료법이 기존치료법보다 우수하다

② 표본자료 수집 및 적절한 통계량 계산

- 가설검정을 위한 실험 및 조사로 관심 대상 변수의 자료를 수집하고, 모수 추정을 위한 통계량(추정량) 계산
- 예) 두 집단의 모평균 차이를 검정하는 경우 : 두 집단의 표본평균과 표본분산을 계산

  (귀무가설 : 두 집단의 모평균은 동일하다)

③ 검정통계량(test statistic) 계산

- 귀무가설을 채택하는 기준 값 제공
- 귀무가설이 참이라는 가정하에 검정에 필요한 검정통계량 계산
- 예) 두 집단의 모평균 차이 검정을 위한 검정통계량

  (모분산이 동일하다는 가정)

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)_0}{se\left(\overline{X}_1 - \overline{X}_2\right)} \sim T\left(df = n_1 + n_2 - 2\right)$$

④ 유의수준 (significant level) 설정

- 유의수준=귀무가설이 참임에도 귀무가설을 기각하는 확률의 크기

$$\alpha = \Pr\left(H_0 \text{ reject} \mid H_0 \text{ true}\right) = Type\ \ I\ \ error$$

⑤ 기각역(critical region) 설정 및 판단

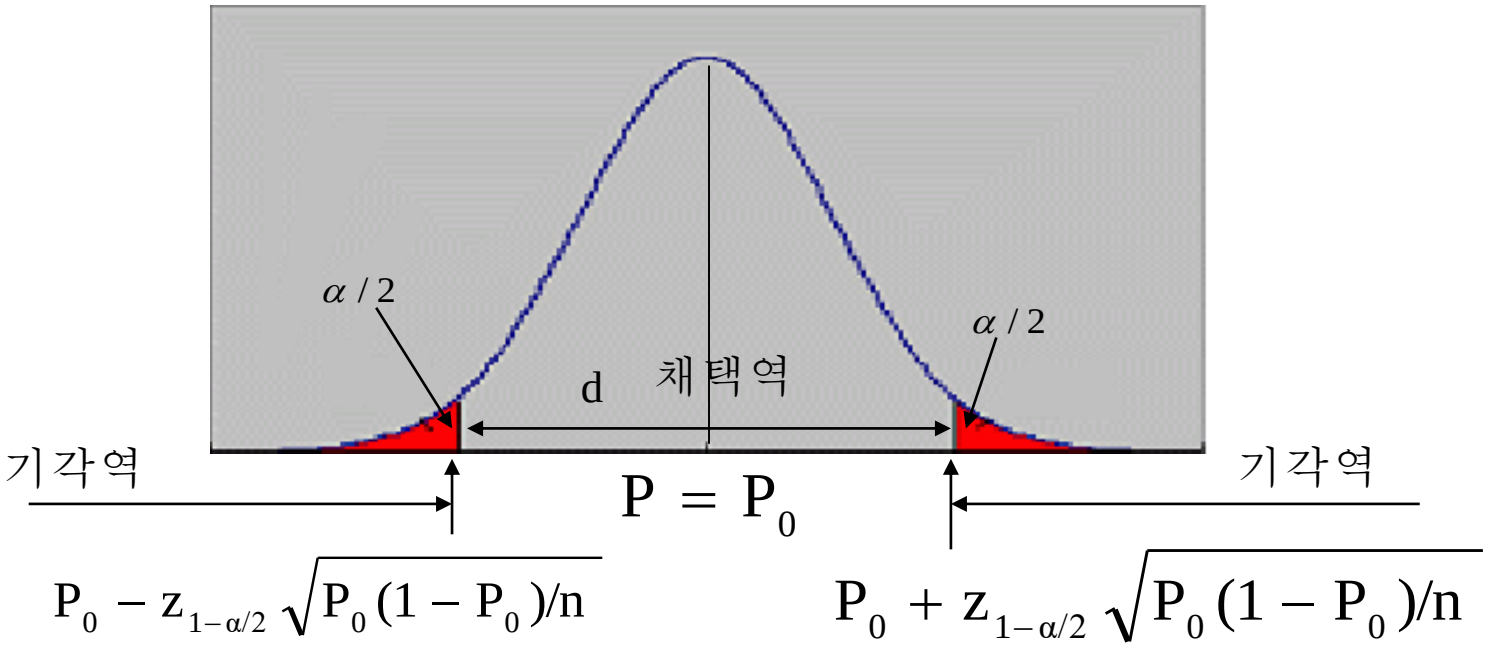
- 기각역 : 귀무가설을 기각하는 영역
  - 임계값, 기각역, 유의확률 등으로 계산
- 유의확률(significant prob. = p-값)
  - 유의수준 (significant level)과 비교
  - p-값 : 귀무가설이 참인 조건에서 표본에서 얻어진 결과로부터 귀무가설을 채택할 확률의 크기

$$p - value = \Pr\left(H_0 \text{ is accepted from sample} \mid H_0 \text{ true}\right)$$

- 기각역 : 대립가설의 형태에 따라 다르게 표현
  (단측 검정, 양측검정)
- 귀무가설 기각 기준

$$\text{if } p - value < \alpha \text{ then } H_0 \text{ is rejected}$$

- **예 : 하나의 모집단에 대한 비율 검정**
  - **가설 :** $H_0 : P = P_0$ **vs** $H_1 : P \neq P_0$

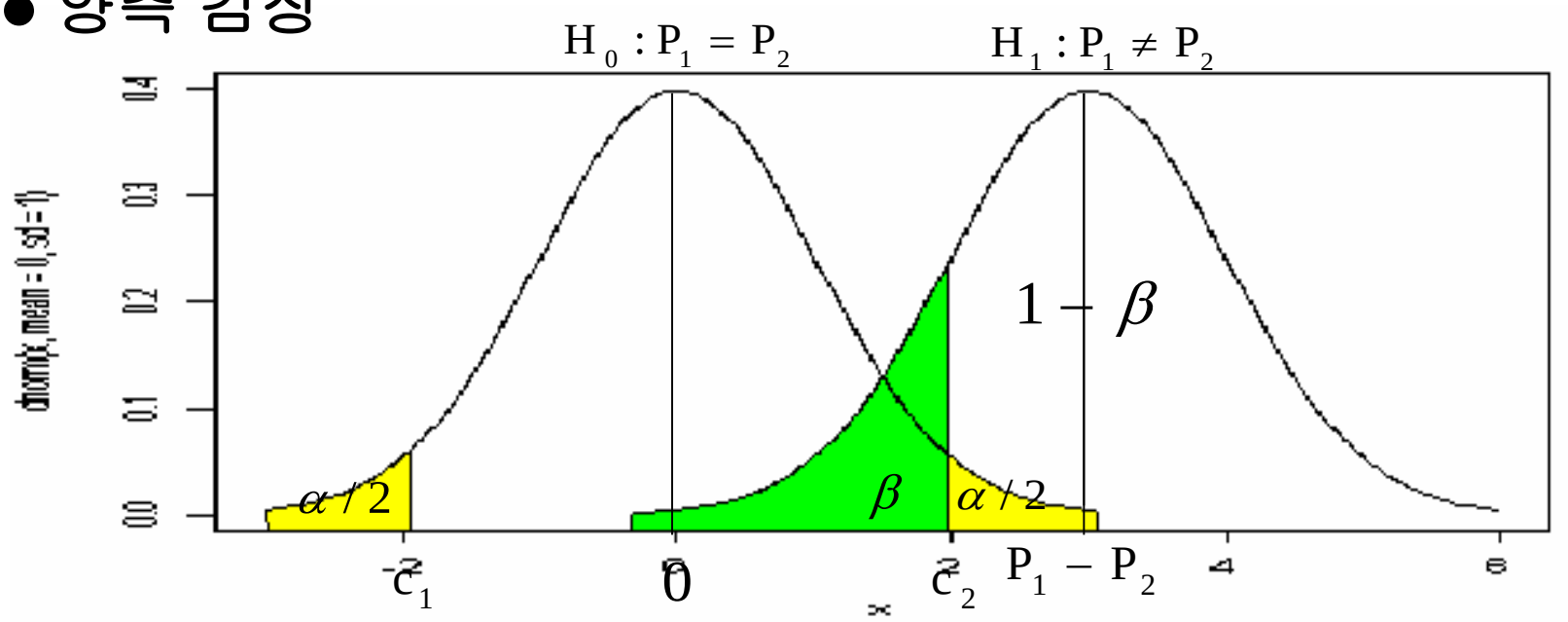


가설검정의 오류

- 통계적 검정에서의 오류
  - 모집단에 대한 가설은 표본을 통해 판단하기 전에는 미지이므로 가설의 참/거짓인 상황에 따라 두 가지 판단오류가 발생
  - 일반적으로 제 1종의 오류를 기준으로 가설검정 수행
  - 의학 연구에서는 제 1종의 오류를 고정시킨 후 제 2종의 오류가 최소가 되도록 기각역을 정한 후 가설검정 수행

사실여부 \ 판단여부		H_0	
		Accept	**Reject**
H_0	**True**	정확한 판단 신뢰수준 $1-\alpha$	제 1종의 오류 유의수준 α
	False	제 2종의 오류 β	정확한 판단 검정력(power) $1-\beta$

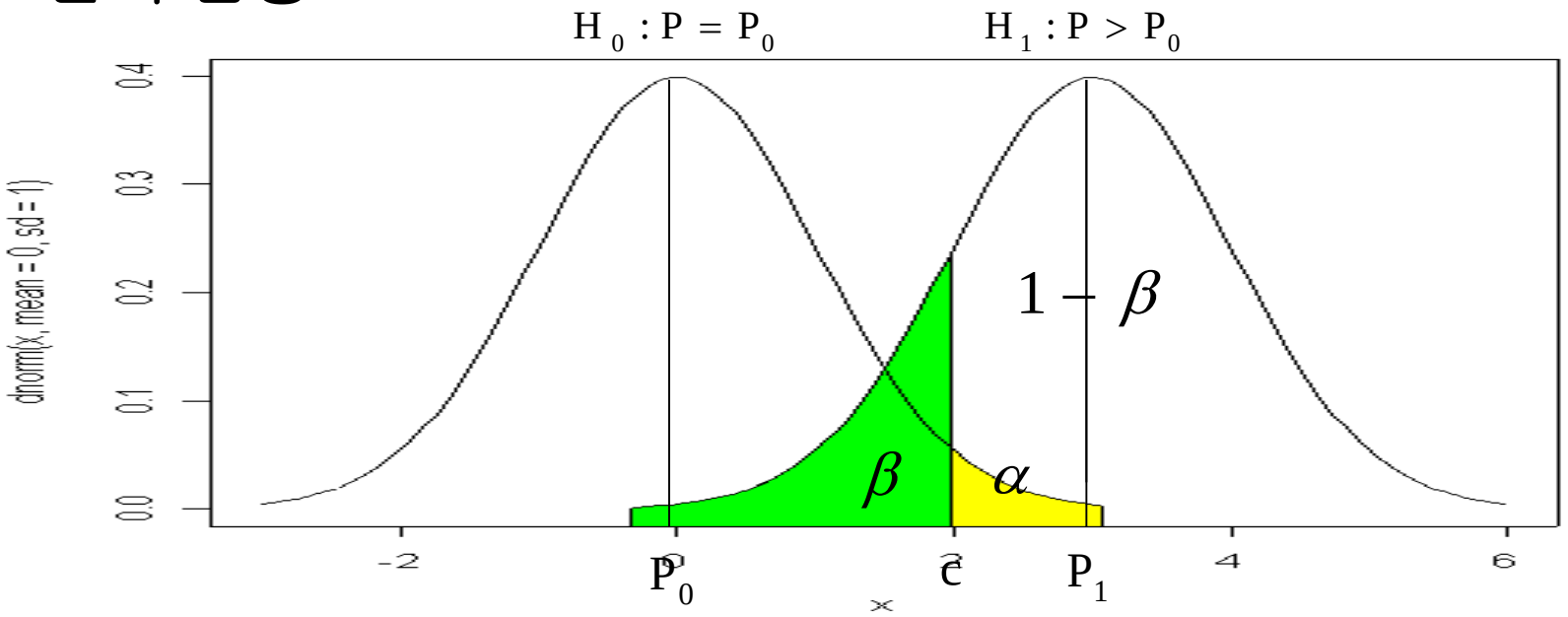
● 양측 검정



$$c_2 = 0 + z_{1-\alpha/2}\sqrt{2\bar{P}(1-\bar{P})/n} \quad \text{under} \quad H_0, \quad \text{where} \quad \bar{P} = (P_1 + P_2)/2$$

$$c_2 = (P_1 - P_2) - z_{1-\beta}\sqrt{(1/n)\left\{P_1(1-P_1) + P_2(1-P_2)\right\}} \quad \text{under} \quad H_1, n_1 = n_2 = n$$

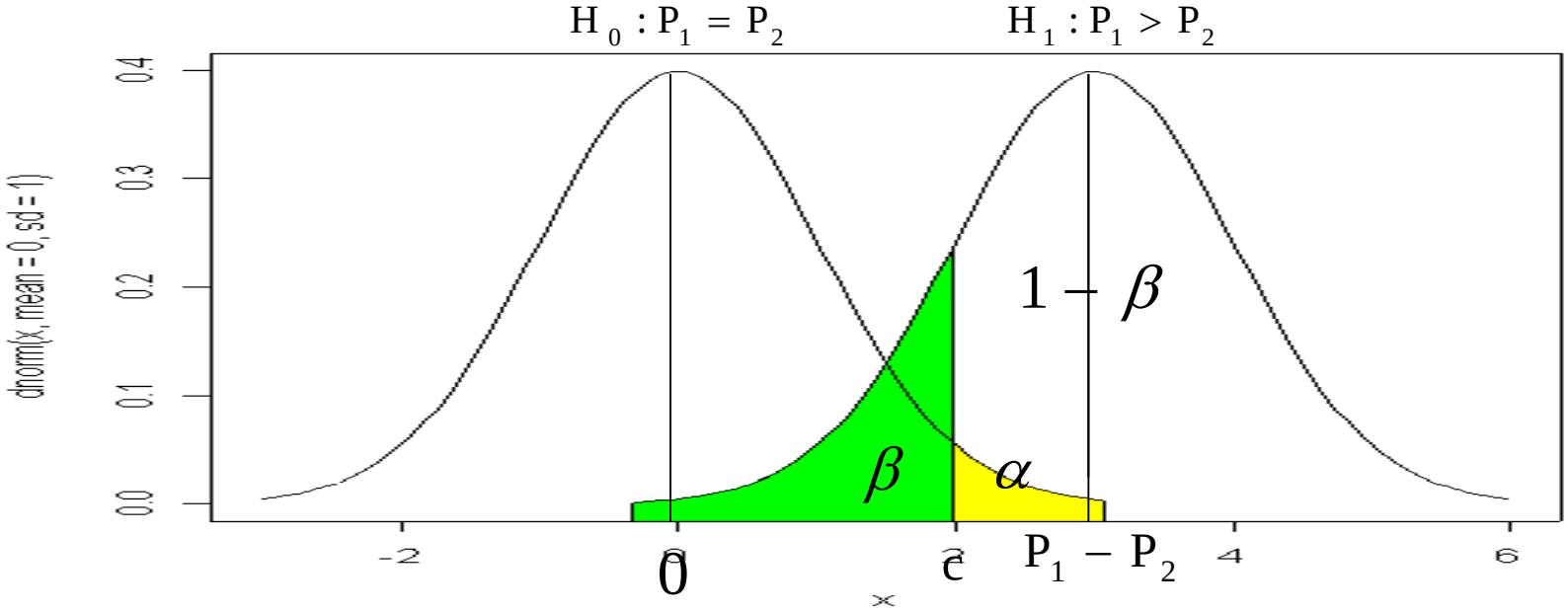
● 단측 검정



$$c = P_0 + z_{1-\alpha}\sqrt{P_0(1-P_0)/n} \quad \text{under} \quad H_0$$

$$c = P_1 - z_{1-\beta}\sqrt{P_1(1-P_1)/n} \quad \text{under} \quad H_1$$

● 단측 검정



$$c = 0 + z_{1-\alpha}\sqrt{2\bar{P}(1-\bar{P})/n} \quad \text{under} \quad H_0, \quad \text{where} \quad \bar{P} = (P_1 + P_2)/2$$

$$c = (P_1 - P_2) - z_{1-\beta}\sqrt{(1/n)\left\{P_1(1-P_1) + P_2(1-P_2)\right\}} \quad \text{under} \quad H_1,$$

assuming $n_1 = n_2 = n$

추측통계학의 분석방법

- 모집단의 특성(분포) 가정에 따라 구분
  - 모수 방법(parametric method) : 모집단 분포에 대한 가정을 전제로 분석하는 방법
  - 비모수 방법(nonparametric method) : 모집단에 대한 가정을 하지 않고 모집단의 형태에 관계없이 수집자료로부터 직접 분석하는 방법
    (분포 무관 방법, distribution-free method)
- 비모수방법으로 분석하는 경우
  - 측정자료의 모집단 분포를 모를 때
  - 측정자료가 특정분포 형태를 따르지 않을 때
  - 자료의 수가 작아 모집단 분포를 확신할 수 없을 때
  - 수집자료가 명목형 혹은 순서형일 때

- 비모수분석 방법의 기본 자료 유형
  - 측정자료를 크기순으로 정렬하여 순위나 순위의 합을 기본 자료로 이용
- 비모수분석의 판단 기준
  - 측정자료의 순위나 순위 합을 이용하여 우연(random)에 의해 발생되는 확률을 계산하여 판단
- 비모수 방법 적용의 추가 사항
  - 자료의 수가 많을 때 적용이 쉽지 않음
  - 집단 비교 시 각 집단의 수가 동일할 때 비모수 방법은 검정력이 낮으므로 모수 방법을 이용하는 것이 바람직

자료분석을 위한 Tip

- **정확한 분석방법을 위한 판단 기준**
  - 분석 목적
  - 분석하고자 하는 변수의 척도
  - 분석방법의 가정

- 예)
  1) 정상군과 환자군의 콜레스테롤의 차이가 있는 가? (비교기준변수 : 집단/명목형, 분석변수 : 콜레스테롤/비율형) (t-검정)
  2) 고혈압군, 중증도군, 정상군의 간기능에 차이가 있는 가? (비교기준변수 : 집단/명목형, 분석변수 : 간기능수치/비율형) (ANOVA)
  3) 암의 형태에 따라 생존여부 및 생존기간은 얼마나 될까? (비교기준변수 : 암의 형태/명목형, 분석변수 : 생존여부 및 생존기간/명목형 및 비율형) (생존분석)

- 대부분 두 변수이상의 관계를 이용해 분석 시도
  - 기본 가정 : 대부분 분석 목표가 되는 변수의 측정자료는 모집단 분포로 정규분포로 가정

- 두 변수 이상의 분석 목적에 대한 검토
  - 상호관련성 분석
    - 두 변수를 동등한 관계로 가정하여 두 변수의 관련성을 분석
    - 인과성 추론을 위한 기본분석으로 널리 이용
      - 상관관계 분석, 독립성을 위한 카이제곱 검정
      - 다변량분석 : 군집분석, 주성분/요인분석 등
  - 인과관계 추론을 가정한 분석
    - 원인이 되는 변수를 독립(혹은 설명)변수
    - 결과로 나타나는 변수를 종속(혹은 반응)변수
    - 대부분의 통계 분석에 해당
      - 집단간 차이 분석 : t-검정, 대응자료 t-검정, ANOVA 등
      - 동일성을 위한 카이제곱검정
      - 회귀분석, 분산분석, 로지스틱회귀분석 등
      - 다변량분석 : 판별분석, 로그선형모형 등

모수방법과 비모수방법

• **모수방법에 대응되는 비모수 방법**

분석 목적			모수 방법	비모수 방법
하나의 집단에 대한 분석 (분석자료 계량형)			z-test , t-test	Kolmogorov-Smirnov (K-S) test, Runs test
상호관계	상관분석		Pearson's corr.	Spearman's corr. Kendall's tau Stuart's tau
인과관계성 추론	명목형 (빈도자료)	두 독립 집단 두 대응 집단 세 집단 이상	$\chi^2 - test$	Fisher's exact test McNemar test Cochran Q test
	계량형	두 독립 집단	indep. t-test	Wilcoxon rank sum test, Mann-Whitney test, Median test
		두 대응 집단	paired t-test	Wilcoxon signed rank test, Sign test
		세 집단 이상 이원배치	ANOVA 2-way ANOVA	Kruscal-Wallis test Friedman's 2-way ANOVA

모수 방법와 비모수 방법의 비교

비모수 방법의 적용

- **변수의 모집단 분포를 모르거나 알려지지 않은 경우**
- **자료의 수가 작아서 분포가 불확실한 경우**
- **명목자료나 순위자료로 측정된 자료 분석에 주로 이용**

분석 목적		모수 방법	비모수방법
적합도 검정		카이제곱 검정 S-W 검정	카이제곱 검정 K-S 검정 이항분포 검정
무작위성 검정		랜덤성 검정	연 검정
분포 동질성 검정/ 자료의 중심값	대응자료	대응 T-검정	부호 검정 윌콕슨 부호순위검정 맥니마 검정
	3집단 이상 (동일표본)	MANOVA	프리드만 검정 켄달 검정 코크란 Q 검정

분석 목적		모수 방법	비모수방법
분포의 동일성 검정/ 자료의 중심값 비교	독립 2집단	T-검정	맨-휘트니 검정 (순위합 검정) K-S 검정 월포비치 검정 중앙값 검정
	3집단 이상 (대응표본)	ANOVA	중앙값 검정 K-W 검정 Jonckheere 검정
상관관계 분석/ 독립성 검정		피어슨 Corr.	스피어만 Corr. 카이제곱 검정 (연관성 검정)

Types of data

Type	scale	properties		example	Numeric type
Qualitative data	Nominal scale	categorical	lowest level	gender, experimental gr., race,	discrete
	Ordinal scale	ordinal	ranking, larger than	edu. Level, response type (−/0/+)	
Quantitative data	Interval scale	Equally interval	non-division, relative zero	temperature	Discrete continuous
	Ratio scale	Numerical data	absolute zero, ±,×/÷	blood pressure, score	

- Other types of data to be encountered in medical research
  - Ranks : the relative position of members of a group
  - Percentages : the ratio of two quantities
  - Rates : to convert an observed frequency to a rate (per 1000)
  - Ratios : the frequency of events to the expected number of events
  - Scores : indirect measurements such as 0,+,++,+++ etc (ordered type)

		척도점에 내포된 가정				비교 방법	대표값	적용 예
		분류	순위	등간격	절대0점			
정성적 (비정량적)자료 nonmetric data	명목 척도	○	×	×	×	확인, 분류	최빈값 Mode	· 성별/혈액형 · 질병분류 · 존재유무
	서열 척도	○	○	×	×	순위 비교	중앙값 median	· 고통정도 · 암증상단계 · 중요도/만족도
정량적 자료 metric data	등간 척도	○	○	○	×	간격 비교	산술평균 mean	· 온도/체온 · 광고인지도 · IQ · 혈압
	비율 척도	○	○	○	○	절대 크기	기하평균 조화평균	· 연령 · 점수 · 무게 · 소득

2. 모수 방법

1. 집단 비교 분석 방법 : 모수 방법

- 가정 :

  ✓반응 결과가 연속형 자료로 수집
  ✓측정자료~정규분포 가정

1-1. 두 독립 집단의 평균 비교

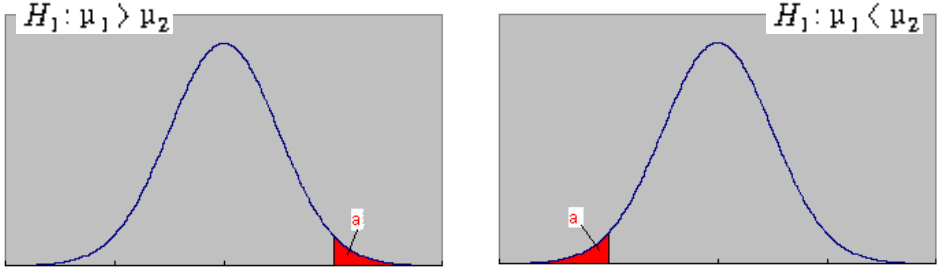
- 정규모집단 : 모분산이 알려진 경우
  - 가정 : 서로 독립인 두 집단의 자료가 정규분포라고 가정 $x_1, x_2, \cdots, x_{n_1} \sim N(\mu_1, \sigma_1^2),\ y_1, y_2, \cdots, y_{n_2} \sim N(\mu_2, \sigma_2^2)$
  - 관심 사항(모수) : $\mu_1 - \mu_2$
  - 추정량(통계량) : $\bar{x} - \bar{y}$
    - 추정량의 표본분포 $\bar{x} \sim N(\mu_1, \sigma_1^2/n_1) \quad \bar{y} \sim N(\mu_2, \sigma_2^2/n_2)$

$$\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2) \quad \text{or} \quad N(\mu_1 - \mu_2, \sigma_0^2\left(1/n_1 + 1/n_2\right))$$

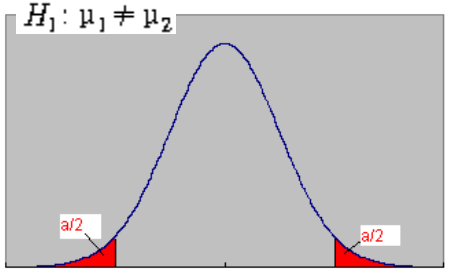
  - 검정통계량 $z_0 = \dfrac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0,1)$ under $H_0 : \mu_1 - \mu_2 = \delta_0$

◦ p-value

• 단측검정 $p - value = \Pr(Z > |z_0|)$



• 양측검정 $p - value = 2 \times \Pr(Z > |z_0|)$



◦ 귀무가설의 기각 기준

• 유의수준 if $\alpha > p - value$ then H_0 is rejected.

정규모집단 : 모분산이 알려지지 않은 경우

- 가정 : 서로 독립인 두 집단의 자료가 정규분포라고 가정

$$x_1, x_2, \cdots, x_{n_1} \sim N(\mu_1, \sigma_1^2),\ y_1, y_2, \cdots, y_{n_2} \sim N(\mu_2, \sigma_2^2)$$

- 관심 사항(모수) : $\mu_1 - \mu_2$
- 추정량(통계량) : $\bar{x} - \bar{y}$
  - 추정량의 표본분포 $\bar{x} \sim N(\mu_1, \sigma_1^2/n_1)$ $\bar{y} \sim N(\mu_2, \sigma_2^2/n_2)$

$$\bar{x} - \bar{y} \sim N(\mu_1 - \mu_2, Var(\bar{x} - \bar{y}))$$

  - 분산추정량

$$\hat{Var}(\bar{x} - \bar{y}) = \begin{cases} s_1^2/n_1 + s_2^2/n_2 & , \ \sigma_1 \neq \sigma_2 \\ s_p^2\left(1/n_1 + 1/n_2\right) & , \ \sigma_1 = \sigma_2 \end{cases}$$

◦ 합동분산 s_p^2 의 추정

$$s_p^2 = \frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{(n_1-1)+(n_2-1)} = \frac{\sum (x_i-\bar{x})^2 + \sum (y_j-\bar{y})^2}{n_1+n_2-2}$$

◦ 자유도(Satterthwaite 자유도) $w_1 = s_1^2/n_1$, $w_2 = s_2^2/n_2$

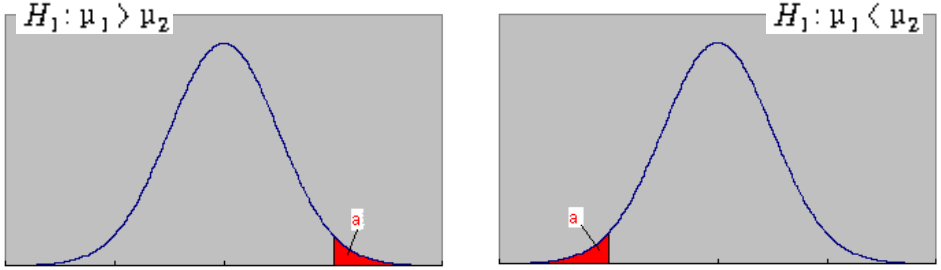
$$df = \begin{cases} (w_1+w_2)^2 \Big/ \left(w_1^2/(n_1-1) + w_2^2/(n_2-1) \right) & , \ \sigma_1 \neq \sigma_2 \\ n_1+n_2-2 & , \ \sigma_1 = \sigma_2 \end{cases}$$

◦ 검정통계량 $t_0 = \dfrac{\bar{x}-\bar{y}-\delta_0}{se(\bar{x}-\bar{y})} \sim T(df)$ under $H_0 : \mu_1-\mu_2 = \delta_0$

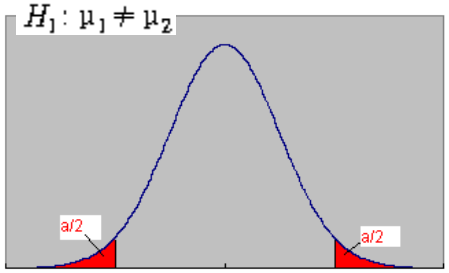
$$z_0 = \frac{\bar{x}-\bar{y}-\delta_0}{se(\bar{x}-\bar{y})} \sim N(0,1) \quad \text{under} \quad n_1,\ n_2 \geq 30$$

◦ p-value

• 단측검정 $p - value = \Pr\left(T\left(df\right) > \left| t_0 \right|\right)$



• 양측검정 $p - value = 2 \times \Pr\left(T\left(df\right) > \left| t_0 \right|\right)$



◦ 귀무가설의 기각 기준

• 유의수준 if $\alpha > p - value$ then H_0 is rejected.

[참고] 모분산의 동일성 검정

- 가정 : 서로 독립인 두 집단의 자료가 정규분포라고 가정

$$x_1, x_2, \cdots, x_{n_1} \sim N(\mu_1, \sigma_1^2),\ y_1, y_2, \cdots, y_{n_2} \sim N(\mu_2, \sigma_2^2)$$

- 관심 사항(모수) :

$$\sigma_1^2 = \sigma_2^2 \quad \text{or} \quad \sigma_1^2 \Big/ \sigma_2^2 = 1$$

- 추정량(통계량) : $\hat{\sigma}_1^2 = s_1^2$ and $\hat{\sigma}_2^2 = s_2^2$
  - 추정량의 표본분포 $s_1^2 \sim \chi^2(n_1 - 1)$ and $s_2^2 \sim \chi^2(n_2 - 1)$
  - 검정통계량 $F_0 = \dfrac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \dfrac{s_1^2 / s_2^2}{\sigma_1^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$
- 기각역

$$H_1 : \sigma_1^2 / \sigma_2^2 > 1 \ \Rightarrow\ F_0 > F_{1-\alpha}(n_1 - 1, n_2 - 1) \quad \text{where} \quad \alpha \ \text{is lower prob..}$$

$$H_1 : \sigma_1^2 / \sigma_2^2 < 1 \ \Rightarrow\ F_0 < F_{\alpha}(n_1 - 1, n_2 - 1)$$

$$H_1 : \sigma_1^2 / \sigma_2^2 \neq 1 \ \Rightarrow\ F_0 < F_{\alpha/2}(n_1 - 1, n_2 - 1) \ \text{or} \ F_0 > F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$$

$$\text{where} \quad F_{\alpha/2}(n_1 - 1, n_2 - 1) = F_{1-\alpha/2}(n_2 - 1, n_1 - 1)$$

1-2. 대응(짝) 집단의 평균 비교

- 대응 자료(paired data) $d_i = x_i - y_i$

표본번호	Before	After	차이
1	x_1	y_1	d_1
2	x_2	y_2	d_2
⋮	⋮	⋮	⋮
n	x_n	y_n	d_n

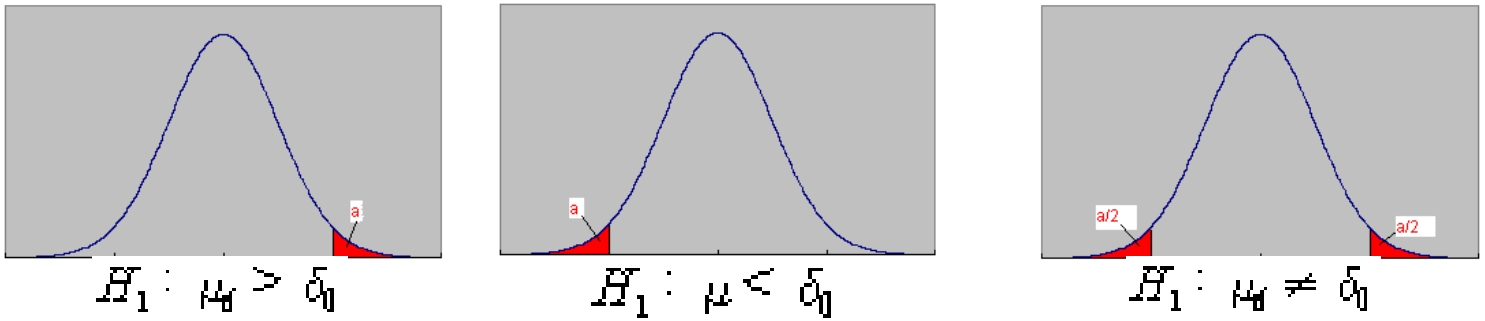
  - 가정 : 서로 대응되는 자료의 차이(한 집단으로 간주)를 정규분포로 가정
    표본간 독립, 표본의 자료는 종속
    $d_1, d_2, \cdots, d_{n_1} \sim N(\mu_d, \sigma_d^2)$
  - 관심 사항(모수) : $\mu_d = \mu_1 - \mu_2$
  - 추정량(통계량)과 표본분포 :

$$\hat{\mu}_d = \bar{d} = \bar{x} - \bar{y} \sim N\left(\mu_d, \sigma_d^2 / n\right)$$

◦ 검정통계량

$$t_0 = \frac{\bar{d} - \delta_0}{s_d / \sqrt{n}} \sim T(df = n-1) \quad \text{under} \quad H_0 : \mu_d = \delta_0$$

◦ 기각역



$$p - value = \Pr\left(T(df) > |t_0|\right) \qquad p - value = 2 \times \Pr\left(T(df) > |t_0|\right)$$

◦ 귀무가설의 기각 기준

• 유의수준 if $\alpha > p - value$ then H_0 is rejected.

1-3. 여러 집단의 평균 비교

- 데이터 형태 : 하나의 변수(요인)인 경우

집단 1	집단 2	• • •	집단 k
y_{11} y_{12} ⋮ y_{1n_1}	y_{21} y_{22} ⋮ y_{2n_2}	• • •	y_{k1} y_{k2} ⋮ y_{kn_k}
n_1 $\bar{y}_1$ s_1	n_2 $\bar{y}_2$ s_2	• • •	n_k $\bar{y}_k$ s_k

$$n = \sum n_i$$

$$\bar{y} = \sum \sum y_{ij} / n$$

  - 가정 : k개 모집단은 서로 독립이며, 정규분포를 가정
    - i번째 모집단은 $N(\mu_i, \sigma^2)$ 이며, n_i 개의 표본을 추출하여 자료 수집한다고 가정
    - k개 모집단의 분산은 모두 동일하다고 가정

- 분산분석모형

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad where \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right)$$

- 변동분해(sum of square, SS)

$$\sum\sum \left(y_{ij} - \bar{y}\right)^2 = \sum\sum \left(\bar{y}_i - \bar{y}\right)^2 + \sum\sum \left(y_{ij} - \bar{y}_i\right)^2$$

$$\Leftrightarrow SST = SSTr + SSE$$

- 자유도 $n - 1 = (k - 1) + (n - k)$

$$\Leftrightarrow df_T = df_{Tr} + df_E$$

- 평균제곱(Mean of SS, MS) $MS_l = SS_l / df_l$, where $l = Tr$ or E

- 검정통계량 $F_0 = MS_{Tr} / MS_E \sim F_{(k-1,\ n-k)}$

ANOVA : 두 요인 이상을 고려하는 경우

- 두 요인 A(i), B(j)의 영향에 대한 연구 : 반복 실험을 가정

	수준B1	수준B2	• • •	수준Bj
수준A1	y_{111} … y_{11n}	y_{121} … y_{12n}	• • •	y_{1J1} … y_{1Jn}
	$\bar{y}_{11.}$	$\bar{y}_{12.}$	• • •	$\bar{y}_{1J.}$
• • •	• • •	• • •	• • •	• • •
수준Ai	y_{I11} … y_{I1n}	y_{I21} … y_{I2n}	• • •	y_{IJ1} … y_{IJn}
	$\bar{y}_{I1.}$	$\bar{y}_{I2.}$	• • •	$\bar{y}_{IJ.}$

$$\bar{y}_{ij.} = \sum_k y_{ijk} / n$$

$$\bar{y}_{i..} = \sum_j \sum_k y_{ijk} /(Jn)$$

$$\bar{y}_{.j.} = \sum_i \sum_k y_{ijk} /(In)$$

$$\bar{y}_{...} = \sum_i \sum_j \sum_k y_{ijk} /(IJn)$$

- 분산분석모형 : α, β 는 주효과, $(\alpha\beta)$ 는 교호작용효과

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \qquad where \quad \varepsilon_{ijk} \sim N(0, \sigma^2)$$

  - 가정 : 오차항~독립, 정규분포, 등분산 가정
  - 효과에 대한 조건 : (모수인자 ; fixed factor)

$$\sum_i \alpha_i = 0, \ \sum_j \beta_j = 0, \ \sum_i (a\beta)_{ij} = \sum_j (a\beta)_{ij} = 0$$

- 분석 관점
  - 주효과(main effects) 에 대한 분석
  - 교호작용효과(interaction effects)에 대한 분석

- 분산분석표

요인	제곱합(SS)	자유도(df)	평균제곱(MS)	F-ratio
A	SSA	dfA=I-1	MSA	MSA/MSE
B	SSB	dfB=J-1	MSB	MSB/MSE
AXB	SSAXB	dfAxB=(I-1)(J-1)	MSAxB	MSAxB/MSE
Error	SSE	dfE=IxJx(n-1)	MSE	
Total	SST	(IxJxn)-1		

  - 여기서 $SST = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$ $SSA = (Jn)\sum_i (\bar{y}_{i..} - \bar{y}_{...})^2$

$SSB = (In)\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$ $SSAxB = n\sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$

$SSE = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$

$MSk = SSk / dfk, \quad k = A, B, AxB$

- 가설의 기각 판단
  - 일반적인 가설 검정의 원리와 동일 : F-분포 이용
  - 검정 과정
    ① 교호작용효과에 대한 검정 : 유의하지 않으면 주효과 검정
    ② 주효과에 대한 검정
    ③ 다중비교 및 사후분석으로 동일집단 분류

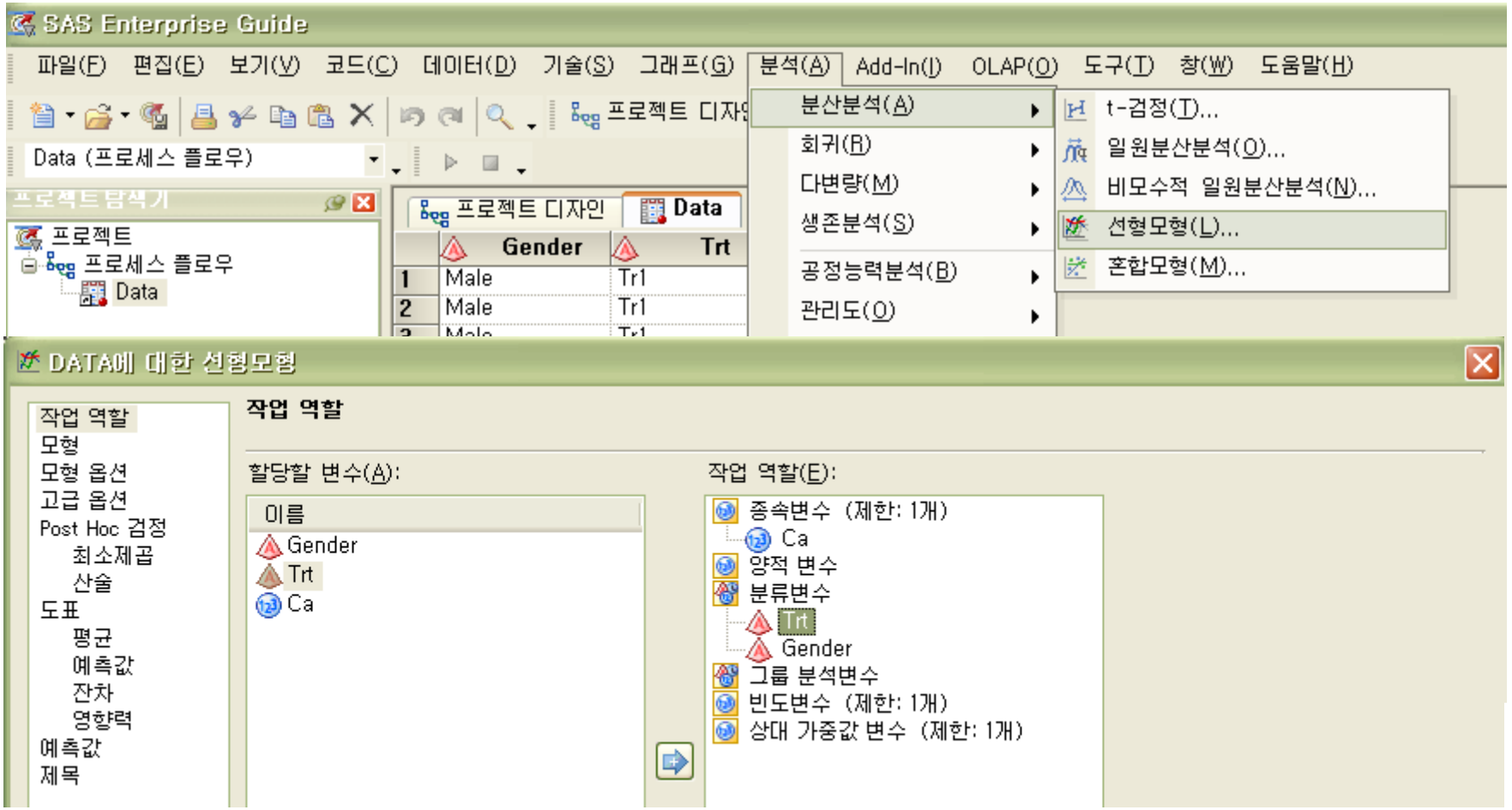
[참고]
1. 분산분석 모형에서 교호작용효과가 유의하면 주효과분석은 의미가 없으므로 교호작용효과가 유의하지 않은 경우에 주효과에 대한 검정을 수행함.
2. 분산분석모형을 표현할 때 교호작용이 유의하면 주효과는 모형에 반드시 포함시켜 표현해야 함.
3. 반복이 없는 경우는 교호작용효과가 존재하지 않음.
4. 사후분석 및 다중비교는 귀무가설이 기각된 경우에 수행하며, 다중비교방법에 따라 집단간 차이유무의 결과가 다르게 나타남.

Example : 혈액 칼슘 자료

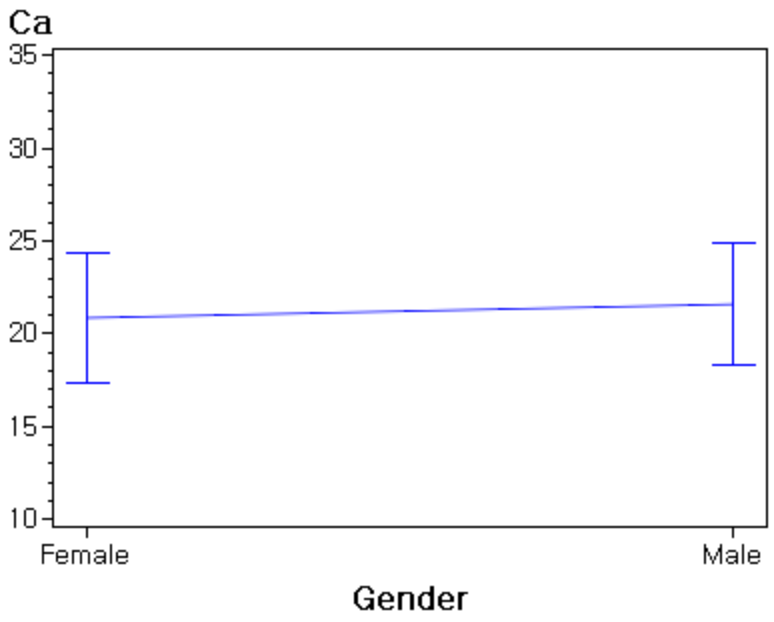
- 호르몬 처리(3종)와 성별(2종)에 따른 혈액 칼슘값의 차이가 존재하는 지에 대한 자료(교재 p.30)
  - 남녀별 15명을 랜덤하게 배정하여 실험

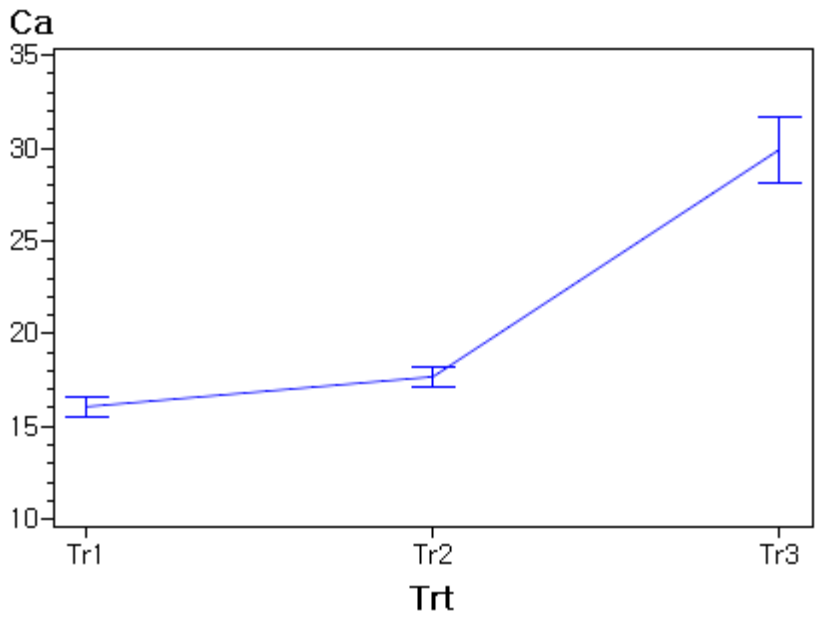
	Tr1	Tr2	Tr3
Male	16.87	19.07	32.05
	16.18	18.77	28.71
	17.12	17.63	34.65
	16.83	16.99	28.79
	17.19	18.04	24.46
Female	15.86	17.20	30.54
	14.92	17.64	32.41
	15.63	17.89	28.97
	15.24	16.78	28.46
	14.80	16.72	29.65

SAS : Two-way ANOVA

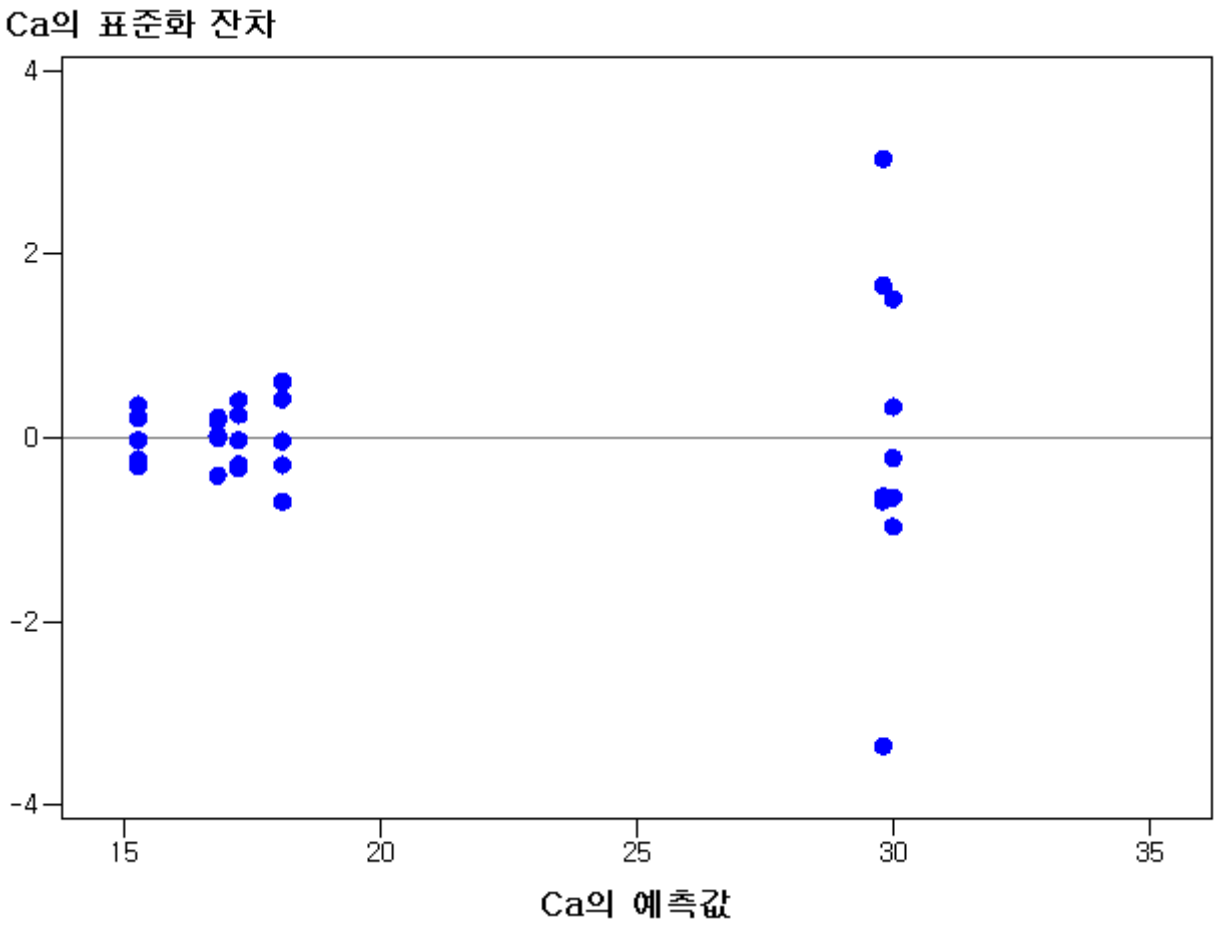



DATA에 대한 선형모형
작업 역할
모형
모형 옵션
고급 옵션
Post Hoc 검정
최소제곱
산술
도표
평균
예측값
모형
분류변수 및 양적변수(V):
Gender
Trt
주효과(M)
교차(O)
지분효과(N)
효과(E):
Gender
Trt
Gender*Trt
DATA에 대한 선형모형
작업 역할
모형
모형 옵션
고급 옵션
Post Hoc 검정
최소제곱
산술
도표
평균
예측값
잔차
영향력
예측값
제목
모형 옵션
가설 검정
절편에 대한 검정 표시(I)
표시할 제곱합
Type I(T)
Type II(Y)
Type III(P)
Type IV(E)
모수 추정값 표시(M)
모수 추정값의 신뢰한계(L)
신뢰수준(V):
95%


Ca
35
30
25
20
15
10
Female
Male
Gender


Ca
35
30
25
20
15
10
Tr1
Tr2
Tr3
Trt

등분산 가정에 위배



- 선형모형 관점의 분석
  - Proc ANOVA
  - Proc GLM

- 분산분석표 : 분산분석모형에 대한 검정

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1154.550147	230.910029	72.64	<.0001
Error	24	76.292400	3.178850		
Corrected Total	29	1230.842547			

- 분산분석모형의 설명력

R-Square	Coeff Var	Root MSE	Ca Mean
0.938016	8.403983	1.782933	21.21533

- 주효과 및 교호작용 효과에 대한 분산분석표

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Gender	1	4.062720	4.062720	1.28	0.2694
Trt	2	1146.642007	573.321003	180.35	<.0001
Gender*Trt	2	3.845420	1.922710	0.60	0.5543

교호작용효과는 유의한 차이(영향)를 보여주지 못하므로 모형은 주효과로만 표현

- 제3 유형의 제곱합(SS3)

효과 통제후 남은 효과를 분석하는 제곱합

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Gender	1	4.062720	4.062720	1.28	0.2694
Trt	2	1146.642007	573.321003	180.35	<.0001
Gender*Trt	2	3.845420	1.922710	0.60	0.5543

- 분산분석모형의 계수 추정치

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	29.81200000	B	0.79735187	37.39	<.0001
Gender Female	0.19400000	B	1.12762582	0.17	0.8648
Gender Male	0.00000000	B	.	.	.
Trt Tr1	-12.97400000	B	1.12762582	-11.51	<.0001
Trt Tr2	-11.71200000	B	1.12762582	-10.39	<.0001
Trt Tr3	0.00000000	B	.	.	.
Gender*Trt Female Tr1	-1.74200000	B	1.59470373	-1.09	0.2855
Gender*Trt Female Tr2	-1.04800000	B	1.59470373	-0.66	0.5173
Gender*Trt Female Tr3	0.00000000	B	.	.	.
Gender*Trt Male Tr1	0.00000000	B	.	.	.
Gender*Trt Male Tr2	0.00000000	B	.	.	.
Gender*Trt Male Tr3	0.00000000	B	.	.	.

교호작용효과의 계수 추정치

[참고] 모형 제곱합에 대한 분할 제곱합의 표현 방법

. Type I SS ; Model문에서 설명요인의 표현 순서에 따라 모형 제곱합이 얼마만큼 증분되는가를 표현
  - 모형에서 요인들의 순서를 어떻게 두느냐에 따라 값이 달라짐

. Type III SS ; 반응변수에 영향을 미치는 여러 요인의 효과를 제거(보정)한 후 각 설명요인이 고유하게 기여하는 제곱합(요인의 고유기여분)을 표현
  - 분산분석모형에서는 동일하나 공분산분석에서는 제 1종 제곱합과 다르게 나타남.

3. 비모수 방법

비모수 방법 : 두 집단 비교

- 위치(location) 관점의 비교
  - 모집단이 동일하면 위치모수(location parameter)가 동일하다는 가정
  - 표본의 위치를 검정
  - 두 집단의 위치 비교 : t-test 와 대응
    - Mann-Whitney 검정(Wilcoxon rank-sum test, WMW 검정) : 두 모집단(분포 동일 가정)의 모평균 차이 유무 비교
  - 다 집단의 위치 비교
    - Kruscal-Wallis 검정 : 위치모수가 다르다는 대립가설, 순위합 검정을 3(>=2이상)집단 이상의 위치 비교로 확대 적용한 방법, 일원배치 분산분석과 대응(다중비교 가능)
    - Jonckheere 검정 : 위치모수가 특정순위의 순위크기로 표현되는 대립가설을 검정하는 방법

- 대응비교의 위치 검정
  - 대응 쌍(matched pair)의 차이에 대해 검정
  - 일 표본 위치 문제로 접근 : paiered t-test 와 대응
  - Wilcoxon signed ranks test
    - 쌍자료 차이에 대한 부호 및 순위를 이용하여 검정
    - 쌍자료 차이의 모집단분포는 “0”을 중심으로 대칭이라는 귀무가설의 타당성 여부를 검정하는 방법

3-1. 독립 표본에서의 비교

2-1-1. Mann-Whitney 검정 : 순위합 검정

- 두 모집단의 정규성, 분산 동일성 가정 불필요
- 두 확률표본의 독립성 가정
- 왜도가 심하거나 특이값이 존재하는 비정규 모집단에 유용
- 두 모집단의 위치모수(평균, 중앙값) 차이 존재 유무 검정
- 동일한 자료(tied data)는 평균 순위 부여
- 가설

$$H_0 : E(X) = E(Y) \text{ or } \Pr(X < Y) = 1/2$$

$$H_1 : E(X) \neq E(Y) \text{ or } \Pr(X < Y) \neq 1/2$$

$$H_1 : E(X) > E(Y) \text{ or } \Pr(X < Y) < 1/2$$

$$H_1 : E(X) < E(Y) \text{ or } \Pr(X < Y) > 1/2$$

- 검정통계량

$$T = \sum R(X_i)$$

$$T_1 = \frac{T - E(T)}{\sqrt{Var(T)}} = \frac{T - n\left(\frac{N+1}{2}\right)}{\sqrt{\frac{n \cdot m}{N(N-1)} \sum_1^N R_i^2 - \frac{n \cdot m (N+1)^2}{4(N-1)}}}$$

- 기각 기준
  - 동점이 없거나 적은 경우 : 검정통계량 분위표 이용

$$T < x_{\alpha/2} \text{ or } \quad T > x_{1-\alpha/2} \Leftrightarrow \; p-value \; \approx 2 \times \Pr(\, Z > \left| z_0 \right|)$$

$$\text{(left)} \quad T < x_{\alpha} \text{ or } \quad \text{(right)} \quad T > x_{1-\alpha} \Leftrightarrow \; p-value \; \approx \Pr(\, Z > \left| z_0 \right|)$$

  - 참고 $\quad x_{1-\alpha} = n\left(N+1\right) - x_{\alpha}$

$n, m > 20$ 일 때 $\quad x_{\alpha} = n\left(N+1\right)\Big/2 + z_{\alpha}\sqrt{nm\left(N+1\right)\Big/12}$

    - p값 계산을 위한 검정통계량의 표준화

$$z_0 \approx \frac{T - E(T)}{\sqrt{nm\left(N+1\right)\Big/12}} \quad \text{where} \quad N = n + m$$

- 기각 기준
  - 동점이 많은 경우

$$T_1 < z_{\alpha/2} \text{ or } \quad T_1 > z_{1-\alpha/2} \Leftrightarrow \; p-value \; \approx 2 \times \Pr(\, Z > \left|T_1\right|)$$

$$\text{(left)} \quad T_1 < z_{\alpha} \text{ or } \quad \text{(right)} \quad T_1 > z_{1-\alpha} \Leftrightarrow \; p-value \; \approx \Pr(\, Z > \left|T_1\right|)$$

2-1-2. Kruscal-Wallis 검정

- 순위합 검정을 3표본 이상의 위치 비교로 확대 적용한 방법
- 일원배치의 분산분석 가설을 검정하는 비모수 방법
- k=2인 경우 이 표본 위치 비교와 동일한 결과 제공
- 동일한 자료(tied data)는 평균 순위 부여
- 가설

H_0 : k 모집단의 분포는 모두 동일하다

H_1 : 적어도 하나의 모집단 분포는 다르다

- 검정통계량

$$T_1 = \frac{\left[\sum_1^k \frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right]}{\left[\frac{1}{N-1}\left\{\sum_1^k \sum_1^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right\}\right]}$$

  - 동점이 없거나 적은 경우

$$T = \frac{12}{N(N+1)} \sum_1^k \frac{R_i^2}{n_i} - 3(N+1)$$

- 기각 기준
  - k=3이면서 각 집단의 표본이 5이하, 동점이 없거나 적은 경우 Kruscal-Wallis검정통계량 분위표 이용 $T > x_{1-\alpha}$ 이면 기각
  - 그렇지 않은 경우
    - 검정통계량 T_1 은 자유도가 k-1인 카이제곱분포로 점근

$$T_1 > \chi^2_{1-\alpha,k-1} \Leftrightarrow p-value \approx \Pr\left(\chi^2_{k-1} > T_1\right)$$

- 다중비교 : 모집단이 다른 분포를 갖는다는 조건

$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| > t_{1-\alpha/2,N-k}\sqrt{S^2\left(\frac{N-1-T}{N-k}\right)}\cdot\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

$$\text{where} \quad s^2 = \frac{1}{N-1}\left\{\sum_1^k\sum_1^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right\}$$

3-2. 대응자료에서의 비교

2-2-1. Wilcoxon 부호 순위 검정

- 쌍자료간 차이의 부호 및 순위를 이용
- 차이의 모집단 분포는 "0"를 중심으로 대칭이라고 가정
- paired t-test 의 비모수 검정
- 동일한 자료(tied data)는 평균 순위 부여
- 차이의 절대값 순위자료에 차이의 부호 부여 $R_i = \text{sgn}\left(D_i\right) \cdot R\left(\left|D_i\right|\right)$
- 차이가 "0"인 자료는 제외한 자료로 분석
- 가설 $H_0 : E(X) = E(Y) \text{ or } M_D = 0$

  $H_1 : E(X) \neq E(Y) \text{ or } M_D \neq 0$

  $H_1 : E(X) > E(Y) \text{ or } M_D > 0$

  $H_1 : E(X) < E(Y) \text{ or } M_D < 0$

- 검정통계량 $T_1 = \dfrac{\sum R_i}{\sqrt{\sum R_i^2}}$
  - 동점이 없는 경우 $T_1 = \dfrac{\sum R_i}{\sqrt{n(n+1)(2n+1)/6}}$
  - $D_i > 0$ 를 기준으로 계산하는 경우

$$T^+ = \sum_1^n \psi\left(D_i\right) \cdot R_i^+ \quad \text{where} \quad \psi\left(D_i\right) = 1, D_i > 0 \text{ and } 0, D_i \leq 0$$

- 기각 기준
  - T^+ 를 이용하는 경우 : 검정통계량 분위표 이용

$$T^+ < x_{\alpha/2} \text{ or } \quad T^+ > x_{1-\alpha/2}$$

$$\text{(left)} \quad T^+ < x_{\alpha} \text{ or } \quad \text{(right)} \quad T^+ > x_{1-\alpha}$$

  - 일반형, 동점이 없거나 적은 경우

$$T_1 < z_{\alpha/2} \text{ or } \quad T_1 > z_{1-\alpha/2} \Leftrightarrow \; p-value \approx 2 \times \Pr(Z > |T_1|)$$

$$\text{(left)} \quad T_1 < z_{\alpha} \text{ or } \quad \text{(right)} \quad T_1 > z_{1-\alpha} \Leftrightarrow \; p-value \approx \Pr(Z > |T_1|)$$

2-2-2 Non-parametric 2-way ANOVA

- Friedman's 2-way ANOVA
  - Assumptions of 2-way ANOVA in parametric method
    - No missing observations per cells
    - No requirement for the data to be Normally distributed, the residual are expected to have a Normal distribution
      (the model is appropriate does not necessarily follow)
  - Non-parametric 2-way ANOVA
    - no fulfil the assumptions of the parametric method
    - the model will not fit well
    - wide variation in the s.d.'s

Friedman's 2-way ANOVA

- Procedures
  - give the rank for each subject
  - compute the rank-sum in the *i*-th group
    (a similar way to the Kruscal-Wallis)
  - Calculate the statistic ***H***
    - *k* groups and *n* subjects (no replications)

$$H = \frac{12}{nk(k+1)} \sum_1^k \left[R_i - n(k+1)/2 \right]^2 \sim \chi^2(k-1) \quad under \quad H_0$$

  - $n(k+1)/2$ is the expected value for R_i if the null hypothesis is true and all groups are the same

$$H = \frac{12}{nk(k+1)} \sum_1^k R_i^2 - 3n(k+1)$$

only one observation in each cell and a few ties

Example : Immersion suit leakage

- Immersion suit leakage(g) during simulated helicopter underwater escape (Light et al, 1987)

	Diving Suit type			
subjects	A	B	C	D
1	308	132	454	64
2	102	526	0	28
3	182	134	96	30
4	268	324	264	90
5	166	228	134	34
6	332	296	458	6
7	198	350	200	90
8	28	274	16	24
Mean	198	283	202.75	45.75
SD	103.06	127.33	178.94	31.63

	Diving Suit type_rank			
subjects	A	B	C	D
1	3	2	4	1
2	3	4	1	2
3	4	3	2	1
4	3	4	2	1
5	3	4	2	1
6	3	2	4	1
7	2	4	3	1
8	3	4	1	2
sum	24	27	19	10
mean	3.00	3.38	2.38	1.25

$$H = \frac{12}{nk(k+1)} \sum_{1}^{k} R_i^{\,2} - 3n(k+1) = 12.45$$

$$p - value = \Pr\left(\chi^2(3) < 12.45\right) < 0.01$$

- Multiple comparisons
  - Comparison of pairs of groups : Wilcoxon matched pair test

- the Friedman's test with two groups is equivalent to an extension of the sign test rather than the Wilcoxon test